Bishop Conrad Senior Secondary School Cantt, Bareilly

BIOTECHNOLOGY PROJECT

On

GENOMICS AND BIOINFORMATICS

Submitted to

Mrs deepali gupta

(PGT biotechnology)

Submitted by

Harsh saxena

Class 12-B

Roll no- 14

CERTIFICATE OF APPROVAL

This is to certify that Harsh Saxena of class XII-B, BISHOP CONRAD SENIOR SECONDARY SCHOOL, BAREILLY has been allotted the topic Genomics And Bioinformatics for Biotechnology project work is partial fulfilment of the A.I.S.S.C.E 2017-2018 Biotechnology practical exam.

_____ _____

Signature of teacher *Signature of principal*

(Mrs DEEPALI GUPTA Mam) *(FR.HEROLD D'CUNHA)*

(PGT Biotechnology).

ACKNOWLEDGEMENT

In the accomplisment of this project sucessfully, many people have best owned upon me their blessing and the heart pledge support, this time i am utilizing to thank all the people who have been concerned with project

Primarily i would like to thank god for bieng able to complete this project with sucess. Then i would like to thank my principal

FR.HEROLD D'CUNHA and biotechnlogy teacher

Mrs Deepali Gupta , whose valuable guidance has been the ones that helped me patch this project and make it full proof sucess his suggestions and his instructions surved as the major contibution twoards the completion of project.

Then i would like to thanks my parent and friends who have helped me with ther valuable suggestions and guidance has been helpful in various phases of the completion of the project .

Last but not the least i wou;ld like to thank my classmates who have helped me a lot.

INTRODUCTION.

The terms "genomics" was coined in 1986 by Thomas roder, to describe the scientific discipline of mapping, sequencing. H.winkler in 1920 had coined the term genome to implicate the complete set of chromosomal and extrachromosomal genes of an organism , a cell , an organelle or a virus.

The field of genomics relies upon bioinformatics, which is the management and analysis of biological information stored in databases. During the mid-1980s to late 1980s, researchers started to use computers as central sequence repository, from where the could be accessed remotely. Later in the early 1990s , genomics was transformed from an academic undertaking to significant commercial endeavour, a course followed by bioinformatics a few years later.

GENOMICS.

In 1920 h.winkler coined the term genome to describe 'the complete set of chromosomal and extra-chromosomal genes of an organism or a virus. In eukaryotes DNA is also present in mitochondriaand chloroplast(in plants only).after a gap of 66 years , Thomas Roderick in 1986 first use the term genomics and describe its scientific discipline of mapping, sequencing and analysing the genome. Now the no number, location, size and organization of all genes required to make up an organism can be known.

GENOME SEQUENCING PROJECTS.

Throughout the world, scientist were trying to sequence the genome completely of several organism of important groups. The reason of sequencing the genome are given below;

- *It provides the knowledge of total no of all the genes*

- *It shows relationship between genes*
- *It provides opportunities to exploit the sequence for desired experimentation*
- *It provides all genetic information about the organism*

Research work on Human Genome Project was successfully completed only due to international collaboration involving 60 countries , 20 genome research centres and more than 1,000 scientists. Upto 1990, some laboratories sequenced 1,00,000nucleotides. Human Genome Project got assisted internationally and significant informations were gathered. On june 26, 2000 the 'working draft' of human genome project was completed. Human genome is large consisting of 3×10^9 *base pairs and a lot of repeated sequence.*

In 1995, the completely sequenced genome of first two smallest bacteria ***haemophilus influenza Rd*** *and* ***Mycoplasma genitalicum,*** *were reported . in 1996 first year* ***Saccharomyces cerevisiae*** *genome sequencing was completed. In 1997 sequencing of genome of the two best studied bacteria* ***Escherichia coli*** *and* ***Bacillius subtilis*** *was completed.*

Methods of gene sequencing

There are several methods used for small scale sequencing of genome. But these methods do not sequence the entire genome. The two direct method used for genome sequencing and one indirect method (using mRNA but not DNA)have been discussed in this session.

1. ***Direct Sequencing Of Bacterial Artificial Chromosome(BAC).***

BAC vetors are stable and introduce complex foreign DNA of 80-100 kb in E.coli cells. Therefore BAC is used in construction of genomic library. Screening of genomic library is done through searching of common restriction fragments. The BAC clone mapping id done just to determine the arrays of contigs (i.e. contigious clones) which overlap. The large DNA fragments are broken into small pieces and the mapped contigs are sequenced. Thus the direct sequencing procedure involves the sequencing of small pieces of DNA taken from adjacent stretches of a chromosome.

2. Random Shotgun Sequencing

Random shotgun sequencing is one approach to sequence genomic DNA. Genomic DNA macromolecules are very long and they contain many genes and other sequences require to build the whole organism. Even with the best of sequencing techniques we get a maximum of 700 bases of sequence information from one single run of experiment. Therefore we need a strategy to sequence the whole DNA. The random shotgun sequencing approach follows a very well known common theme "divide a big problem into small tasks. Solve these small tasks individually. Finally add up all these solution to get final solution". Big genomic DNA molecules are broken down into small fragments, which are cloned in small (2.0kb) and medium (10kb) plasmid vectors. Plasmids have specific sites where these molecules can be inserted through enzymatic procedure. Thus a library is constructed. Now each clone is picked up randomly and sequenced from both ends. By picking many clones and sequencing them, we get large amounts of sequences. Observations show that several of these sequences are identical, some are similar to each other in parts – called ***overlapping parts*** *– whereas few may be just unique. After we feed all these data into a computer program, these sequences are*

joned by finding overlapping parts. The result is we get log pices of DNA sequences. This process of assembling continues until all overlapping parts are exhausted. Finally, we would get a large portion of genomic DNA sequence.

The Expressed Sequence Tag (EST) Approach

The EST approach was pioneered by J.Craig Venter and co-workers at the National Institute of Health(NIH) USA in early 1990s . he developed a new method of investigating the genes by focussing the attention on active portion of the genome as m RNA. Venter and co-workers isolated mRNA molecule (instead of fragments of genomic DNA) and constructed cDNA molecules. They treated DNA as a part of chromosomal DNA and sequenced to create expressed sequence tags (EST). The ESTs were used as handless for isolating the complete genes. Following EST strategy plenty of databases of nucleotide sequence were generated. Consequently it helped to prepare the transcript map of human genome at preliminary level. The EST technique demonstrated the possibility of sequencing all genes to highest levels. This attempt boosted up the growth of genomic industry.

GENE PREDICTION AND GENE COUNTING

Gene prediction is an important problem for computional biology and there are various algorithm that do gene prediction using known genes as a training data set. Since most of the knowledge to carry out these predicts comes from experimentally identified genes this becomes a limitation. Even if we know where the genes are in the genome, it is not entirely clear how to count them. Due to the existence of overlapping genes and splice variants it is difficult to

define the parts of the DNA that should be regarded as the same or several different genes. Nevertheless, for practice purpose (allowing for some 'experimental error') we can count the number of genes in an organism.

Organisms	Genome size (bp)	Number of predicted genes	Part of genome encoding proteins
Escherichia coli	5×10^5	5,000	90%
Saccharomyces cerevisiae	$14x10^6$	6,000	70%
Drosophila melanogaster	$3x10^8$	14,000	20%
Caenorhabditis elegans	$1x10^8$	18,000	27%
Arabidopsis thaliana	$1.25x10^8$	25,000	20%
Homo sapiens	~ $3x10^9$	~ 30,000	<5%

FUNGCTIONAL GENOMICS

Fungctionqal genomics dissects the knowledge about genomes to understand the genes and their product functional and interaction. Two exciting new developments are now enabling scientists to get a wealth of clues of this complicated story. The new technique called microarray technology and proteomics provide snapshot of all the genes expressed in a cell or tissue under different environmental conditions. The DNA microarray technology is used for analysing the erxpression for thousands of messenger RNA molecules.

Fluorescence in situ hybridization.

Nick translation technique was first developed by Rigby and Paul Berg in 1977. Using this technique colours can be incorporated into DNA sequences. An enzyme DNA polymerase I is used in this technique. This enzyme perform a host of clean-up function during replication from 3'-> 5' exocatalytic activity

It is a nick translation DNA (or RNA) strand paired to DNA template is simultaneously degraded by the 5'-3' polymerase activity of the enzyme. Hence this enzyme has a role both in DNA repair and removal of RNA primers during replication. If nick is not present, DNA polymerase I and DNase I are added to a bufferd solution containing dNTPs (where nucleotide is labelled with green or red fluorescence dye). DAase I make a nick and hydrogen bonds between nucleotides of two template are now broken in 5'-3' exocatalytic activity and repaired by 5'-3' poly-merase activity. Nick created by DNase I rexposes 3'OH and 5'PO_4 ends. DNApol I adds dNTP at 3'OH end. The nucleotide labelled with fluorescent dye is added. The newly synthesised DNA strands contains fluorescent nucleotides. After nick translation the size fluorescent DNA strand depends upon conc of DNA pol I and incubation time. Generally the fragment size varies from 300-3000bp.

Application of FISH in detection of chromosomal defects.

One of the chromosome defects arising from translocation is the Philadelphia chromosome. This abnormality is found in bone marrow of 90% patients suffering from chronic myleogenous leukemia (CML). taking the blood sample karyotype analysis of lymphocyte preparation was carried out. Reciprocal translocation between the chromosome 9 and 22 CML patients was noticed. Counting such cells that carry out Philadelphia chromosome, one can find out how severe this disease is ?but this method is time consuming.

FISH technology has made easy to detect such defective chromosomes. Scientist identified and isolated the clone which possessed genes associated with CML lymphocytes. The probes were prepared by labelling specific region of chromosome 9 by red colour and that chromosome 22 by green colour following nick translation

method. Smear of lymphocyte cells are prepared and then hybridised with two probes in situ. when hybridised smear is observed under fluorescent microscope, the affected cells appear yellow (after hybridisation mixing of green and red colours impart yellow colour) and the unaffected normal cells appear red and green. Besides, the FISH technology is also useful in detecting the status of a disease during interphase of cell division. By counting the yellow coloured cells status disease can be found out. Similarly effect of chemotherapeutic drug can be assessed by counting the number of yellow coloured cells.

DNA microarray technique.

A major technological advancement was made in the field of molecular biology during the mid 1990s when DNA chips were produced. Ti attracted the interest among the biologist throughout the world. DNA chips are high density miniaturised microarrays of large number of DNA sequences which are attached in a fixed (spotted) location in a systematic order on a solid support e.g. glass plates, slides or nylon. The principle of DNA microarray lies on the base pairing or hybridisation between the nucleotides. Using this technology the presence of one genomic or cDNA sequence in 1,00,000 or more sequences can be screened in a single hybridization. The DNA chips contain known oligonucleotides (20-mers) sequences or cDNA of known function. Thus a sinle DNA chp can give the complete picture of whole genome of an organism. For application in DNA sequencing, DNA chips will have to posses every possible oligonucleotide sequence. Because the maximum sequence read' possible is the square root of the number of oligonucleotide sequences on the chip.

BIOINFORMATICS

(IN SILICO BIOLOGY)

In 21st century biology is being transformed from a purely laboratory-based science to an information science too. The information refers to comprehensive views of DNA sequences, RNA expression and proteins interactions. Due to explosion of sequence and structure information available to researchers, people have become optimistic to get answer of fundamental bio medical problems.

Translation of billions of characters in DNA sequences that make the genome into biologically meaningful information has given birth to a new field of science called bio-informatics.

A more precise definition of bioinformatics is the application of information sciences (mathematics, statistics, and computer science) to increase our understanding of biology. Thus bioinformatics is a multidisciplinary science which aims to use the benefits of computer technologies in understanding the biology of life. Now, as a subject bioinformatics consist of three core areas:

1. *Molecular biology database*
2. *Sequence comparison and sequence analysis*
3. *The emerging technology of microarrays,*

In brief bioinformatics is the management and analysis of biological Information stored in database

DATABASE

A database is repository of sequences (DNA or amino acids) which provides a centralised and homogenous view of its contents. The repository is created and modified through a database management

system (DBMS). Every data item in the database is structured according to a scheme , defined as a set of pre- specified rules through the data definition language. The contents of database can be accessed through a graphical user interface (GUI) that allows browsing through the contents of the repository very much similar as one may browse through the books in library.

SEQUENCES AND NOMENCLATURE

As mentioned earlier that the sequences of digtal symbols are the transformed biopolymers. Indirectly the sequence data means the structure of biopolymer, and structure express the function itshows a reductionist approach. Therefore, the sequence data can ber used as context free.

THE IUPAC SYMBOLS

The International Union of Pure and Applied Chemistry (IUPAC) has made certain recommendations. The nomenclature system in bioinformatics is based on these recommendations

- *Different laboratories of the world follow nomenclature system of IUPAC so that their data set can uniformly and easily be compared.*
- *For rapid reproducibility and uniformity, the database institution and editors(who publish journals and research findings) also follow the recommendation of IUPAC.*

For routine work, the basic IUPAC nomenclature system of nucleic acids and proteins has been discussed in this section. Language used in bioinformatics is given below

The following language is used in bioinformatics:

Alphabets	*=>*	*Nucleotides*
Words	*=>*	*Gene (prokaryotes)*
Sentence	*=>*	*Operon (prokaryotes)*

NOMENCLATURE OF DNA SEQUENCES

It is obvious that nucleotides are the building blocks of DNA, and the nucleotides are constituted by four bases (A, G, T and C) . symbols of these four bases and bases and basis of their nomenclature are used as much as they are spelt. The above table shows symbols, their meaning and bases of nucleic acid sequences.

Often the identity of the sequences at specific position is not clearly identifiable when the sequence data are experimentally determined. It happens due to problems related to other secondary structures or 'compression' artifacts. In compression secondary structures in DNA fragments causes them to move in the gel so that more than one size of fragments may migrate to the same position.

Generally by repeating the experiment and sequencing the complementary strand, this problem can be solved. However, if ambiguities persist in some cases, the probable possibility can be deduced from the gel reads i.e. forward and reverse readings give data from opposite strands of DNA. They provide information about the relative orientations of the ***read pairs*** *(i.e. pair of reading) from the same template of fragments.*

Table 17.1 Summary of single-letter code recommendations by IUPAC.

Symbol	Meaning	Base(s)
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G or A	PuRine
Y	T or C	PYrimidine
M	A or C	Amino
K	G or T	Keto
S	G or C	Strong (3 Hydrogen bonds)
W	A or T	Weak (2 Hydrogen bonds)
H	A or C or T	Not-G, H follows G in the alphabet
B	G or T or C	Not-A, B follows A in the alphabet
V	G or C or A	Not- T (not-U), V follows U in the alphabet
D	G or A or T	Not-C, D follows C in the alphabet
N	G or A or T or C	Any

The concept of directionality

In the biological system the usual direction in which the DNA and RNA are synthesised is in the 5'-3' direction. This is universal and therefore it is helpful to adopt this fact as a way to collect and store data in the sequence databases. The nucleotide sequence are generally present in databases as they have been submitted or published , subject to some conventions which have been adopted for the database as a whole. The sequence are always listed in 5' to 3'. Bases are numbered bases are numbered sequentially with 1 at 5' end of the sequence . the complementary sequence is described with a 'c' indicated next to position of of the sequence . complementary sequence runs 5'-3' but in the opposite direction to given strand. Only one strand is of the DNA sequence is given in a database entry. The complementary strand will have to be inferred using programmes available in various packages or from various Web sites.

In case of proteins, they are synthesised in the cell from N-terminus to the C-terminus. It is useful to adopt this convention in database entry for protein sequencs.

DIFFRENT TYPES OF SEQUNCES

cDNA- *a large number of sequences deposited in database were determined from cDNA molecules. While filling up the sequence entry form you must tick the right position to indicate weather the sequence being deposited is a cDNA sequence. This data will also be provided when a sequence is retrieved.*

Genomic DNA- *sequencing of genomic DNA has become very routine nowadays. The genomic DNA is the store house of information of which expressed part is represented in the cDNA sequences also.*

ESTs – *is an abbreviation for Expressed Sequence Tags. Dr. Craig Venter initiated sequencing of a larger number of cDNA molecules by sequencing one end of each of the randomly picked cDNA clones. Millions of ESTs have been deposited in a special database called dbEST, EST data cDNA clones. EST is used to infer expression patterns by counting the number of ESTs corresponding to each gene divided by the total number of ESTs.*

GST- *in Plasmodium Falcipuram the enzyme Mung Bean Nuclease (MNase) cleaves in between the genes. A genomic DNA library generated by digestion with MNase was used for for gene identification in P.falcipuram. the approach used was similar to ESTs.one read of sequence was obtained from either ends. This data is referred to as genome sequence tag (GSTs). Usually, genomic DNA sequence refers to the nuclear DNA.*

Organelle DNA*- eukaryotic cells have organelles such as mitochondria and chloroplast. These organelles have their own store house of information in the form of organelle DNA. Organelle DNA*

codes for a few genes. The coding information for the rest of the genes resides in the nuclear DNA of the same cell. If an organelle DNA has been sequenced the appropriate position in the sequence submission form must be mentioned .

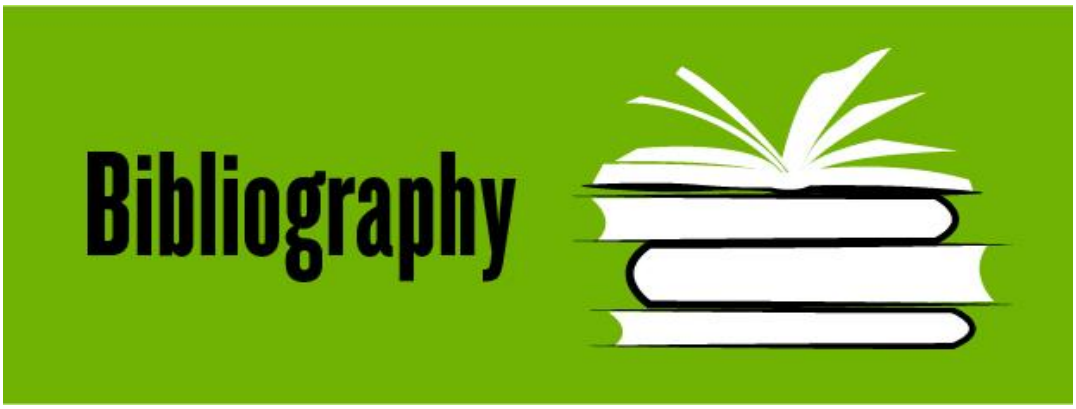
THE BLAST FAMILY OF SEQUENCE-SIMILARITY SEARCH PROGRAMS

The most frequent type of analysis performed on GenBank data is the search for sequences similar to a query sequence. NCBI offers the BLAST family of search programs for this purpose. NCBI's Web interface to the standard BLAST 2.0 program accepts a sequence or accession number as the input query. The search for similarity, performed using an identity matrix for blastn (nucleotide) searches and a PAM or BLOSUM scoring matrix for protein searches, results in a set of gapped alignments, with links to the full document records. Each BLAST alignment is accompanied by an alignment score and a measure of statistical significance, called the Expectation Value, for judging the quality of the alignment. Web BLAST also provides a graphical overview of the alignments, which are color-coded by alignment score and clearly show the extent and quality of the sequence similarities detected by BLAST, as well as the disposition of gaps in the alignments.

The default databases searched by BLAST are the non-redundant (nr) nucleotide and protein databases constructed from the Entrez databases. Several pre-defined specialized databases or subsets may also be searched, and searches may be restricted to sequences from a particular organism. Customized BLAST pages allow a nucleotide query against any combination of 21 complete and 40 incomplete microbial genomes, or against the genomes of malaria-associated pathogens.

Specialized versions of BLAST are also offered to facilitate other approaches to protein similarity searching. Position Specific Iterated BLAST (PSI-BLAST) initially performs a conventional BLAST search to produce alignments from which it constructs a position specific

profile. Subsequent BLAST iterations use this profile matrix in place of the initial query and scoring matrix to find similarities in a database. Pattern Hit Initiated BLAST (PHI-BLAST) takes as input both a peptide query sequence and a peptide pattern, or motif, found within the peptide query sequence. The motif specifies an obligatory match between query and database sequences, about which optimal local alignments are constructed. Another variant, 'BLAST2Sequences' ,can display the similarity between two DNA or peptide sequences by producing a dot-plot representation of the alignments it reports.



INTERNET WEBSITES USED

1. ***https://www.ncbi.nlm.nih.gov***
2. ***https://blast.ncbi.nlm.nih.gov/Blast.cgi***

BOOKS USED

1. ***NCERT OF BIOTECHNOLOGY***
2. ***S.Chand of biotechnology.***